Some Properties of Japanese Sounds through Perceptual Experiments and Spectral Analysis.

AUTHOR(S):

Nakagawa, Seiichi; Sakai, Toshiyuki

STUDIA PHONOLOGICA XI (1977)

Some Properties of Japanese Sounds through Perceptual Experiments and Spectral Analysis

Sei-ichi NAKAGAWA and Toshiyuki SAKAI

Summary

We have been studying on automatic speech recognition and our experience taught ourselves that we had better investigate "What features are important in phoneme recognition?", "Why is phoneme recognition difficult?" and "To which extent is it difficult?" for developing a successful speech recognition system. In this paper, we would try to answer to such questions through perceptual experiments and statistical analyses of spectra.

First, we found the following properties of voiced consonants from perceptual experiments.

(1) The change of speech power (energy) along with time-axis is not the cue of perception of voiced consonants.

(2) More phonemic information of voiced consonant is contained in glide than in a central part of consonant.

(3) Even if the glide is taken out of a VCV utterance, the intelligibility is beyond 60%~70%.

Second, we obtained the following results from statistical analyses of static features of voiced consonants.

(4) The spectral difference between any two vowels is larger than that between two voiced consonants.

(5) The spectral difference between two voiced consonants having the same manner of articulation is particularly small.

(6) /N/, /y/, /w/, /m/, /n/, and /ŋ/ are more influenced by the speaker-factor than by adjacent phonemes (or context).

Third, we tried three-dimensional representation of consonants based on the dynamic features of spectra and investigated the relation among consonants in that space. Finally, we investigated the spectral change of consonants, caused by the factor of speaker and vowel. The results are:

(7) /w/ and /z/ have the property of unvoiced consonants as compared with other voiced consonants.

(8) The spectra of vowels and unvoiced consonants are not subject to the influence of the speaker-factor so much as voiced consonants.

Sei-ichi NAKAGAWA (中川聖一): Assistant, Department of Information Science, Kyoto University.
Toshiyuki SAKAI (坂井利之): Professor, Department of Information Science, Kyoto University.

(9) The spectra of /u/, /i/ and consonants followed /i/ and /u/ are subject to the influence of the speaker-factor more than others.

(10) However, there is no significant difference of speaker-factor among consonants for a long time interval.

I Introduction

Speech science has progressed admirably in speech perception, analysis, synthesis, and recognition. Although we have studied on automatic spoken wird recognition[1] and speech understanding of Japanese sentences[2] for about four years in such environments, we ran into a blank wall. There are two approaches to recognize speech. One is the description of physical features (shape of vocal tract, positions of lip-tongue and jaw, etc.). The other is the statistical processing of observed values (spectrum, formants, zero-crossing number, LPC, etc.). Although the former is promising in future, it has many problems to be solved. The latter, on the other hand, is implicitly related to the physical propetties of sounds, and the processing algorithm is simpler than the former. Therefore, we adopted the latter approach.

Frankly speaking, our experience, however, taught ourselves that we had better investigate in detail "What features are important in phoneme recognition?", "Why is phoneme recognition difficult?" and "To which extent is it difficult?" for developing a successful speech recognition system. In this paper, firstly we point out a few significant points with respect to the automatic recognition of voiced consonants through perceptual experiments. Next, we investigate some characteristics of consonants by using static or dynamic features of short time spectra, and finally we clear some groups of consonants which an automatic recognizer is very difficult to classify.

II Perceptual Experiments of Voiced Consonants

As a means of information media, speech has many kinds of information (linguistics, emotion, personality, sociality, etc.). Therefore, speech is very redundant judging from the viewpoint of linguistic information (or articulation). In order to develop an automatic speech recognizer, we must know where and in what form linguistic information is contained in speech waves. For the purpose of this investigation, there are following three approaches: (a) evaluation of warped speech wave, (b) speech analysis, and (c) speech synthesis[3]. Since (b) and (c) represent or approximate natural speech by using a few feature parameters, we can get easily the evaluation of a feature parameter with respect to linguistic information. On the other hand, (a) is difficult to control the warp, although the obtained results are reliable because there is a direct relationship between natural speech and its processed speech. We adopted the evaluation method by warped speech waves in order to investigate the phonemic information of voiced

consonants. This method is divided into five types:

(1) direct warping of speech wave (ex. zero-crossing wave)[4)~6)].

(2) cutting off from speech wave in time domain[7)~10)].

(3) cutting off from speech wave in frequency domain[11)~13)].

(4) addition of noise[11),12),14)].

(5) exchange or connection of fragments of speech wave[15)~17)].

In this paper, we describe the perceptual experiments of voiced consonants by using methods of (1) and (2).

II–I Experimental System and Test Material

A male adult (Mr. Ukita) uttered 165 consonants in vowel environments, that is, VCV syllables (V=vowel, C=voiced consonant). These materials were sampled by 10KHz, digitized to 10-bits and stored in a disk file of the medium-sized computer NEAC 2200/250 through the in-house computer network KUIPNET[18)] while controlling by the mini-computer MELCOM/70. The stored speech was transformed into a warping wave by MELCOM/70 and into an analog wave by D/A converter. It was passed into a low pass filter (cutting frequency=4.2 KHz) and recorded at analog tapes. The subjects of listening test were five male adults. First, we recorded the speech without warping in order to check the performance of this system, i.e., A/D, D/A, filter and so on. Table 1 shows this confusion matrix for voiced consonants. The intelligibility was about 97.7%. There was no confusion for vowels (This fact will be holded for following experiments). Table 2 represents the confusion matrix of clipped speech for our information.

Table 1. Playback (A/D→D/A).

(97.7%)

in \ out	m	n	g̃	b	d	r	z
m	98.8	1.2					
n		100					
g̃			100				
b				100			
d				2.5	97.5		
r					1.2	98.8	
z					11.1	1.4	87.5

Table 2. Zero-crossing wave.

(63.8%)

in \ out	m	n	g̃	b	d	r	z
m	63.0	25.9	3.7	4.9		2.5	
n	7.4	72.8	3.7	1.2	2.5	11.1	1.2
g̃		3.7	70.4	3.7	21.0	1.2	
b	1.2	3.7	4.9	70.7	16.0	1.2	2.5
d	1.2	8.6	14.8	7.4	60.5	1.2	6.2
r		8.6	7.4		4.9	77.8	1.2
z		6.9	20.8	2.8	40.3	1.4	27.8

II-2 Perceptual Experiment by Normalization of Speech Power

The level of speech power does not depend on only the subglottic pressure, but also depends on the shape of vocal tract. In general, the power of voiced consonants is smaller than that of vowels. Furthermore, the change of power depends on a kind of consonants and context. We examine whether the manner of change gives the cue for perception of voiced consonants or not.

The speech power at a time interval $[t, t+T-1]$ is defined as follows:

$$E_{t,t+T-1}=\left(\frac{1}{T}\sum_{i=t}^{t+T-1} s^2(i)\right)^{1/2},$$

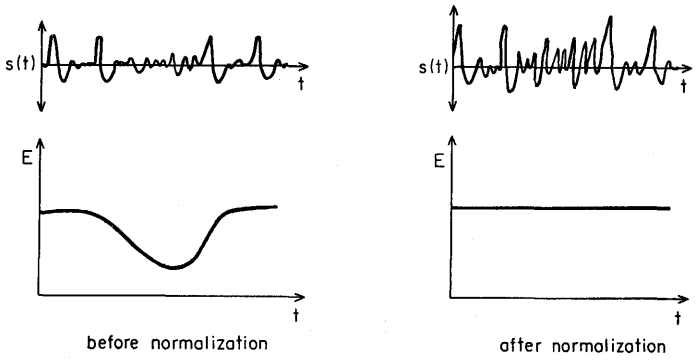

Fig. 1. Normalization of speech power.

Table 3. Normalization of speech power.

(96.4%)

in \ out	m	n	g̃	b	d	r	z
m	1.2	97.5	1.2				
n		100					
g̃			100				
b	2.5		4.9	92.6			
d		2.5		2.5	92.6		1.2
r						100	
z		1.4	1.4		1.4	4.2	91.7

where s(t) is the amplitude of speech wave at time t, and T is fixed interval. We set T to 10 ms. Speech wave, s(t), is normalized by the power as follows:

$$\overline{s(t)}=\frac{s(t)}{E_{t,t+T-1}},\ \overline{s(t+1)}=\frac{s(t+1)}{E_{t,t+T-1}},\ \cdots,\ \overline{s(t+T-1)}=\frac{s(t+T-1)}{E_{t,t+T-1}}$$

Such eliminates the change of speech power along with time-axis. Fig. 1 illustrates this processing. Table 3 shows the confusion matrix of this listening test. From this result, we find that the change of speech power is not the cue for perception of voiced consonants, although many automatic speech recognition systems take advantage of the change for the detection of voiced consonants.

II-3 Role of Glide for Perception of Voiced Consonants

Voiced consonants are transient sounds except for nasal. In particular, there is a significant transient part between a consonant and its preceding vowel or following vowel. We call them as on-glide or off-glide, respectively. The glide contains the influence of coarticulation by a consonant and vowel. If we can identify a voiced consonant in the case of eliminating the glide, an automatic recognizer will not require many standard patterns for voiced consonant recognition. In this section, we investigate which contains more linguistic information, stationary or glide.

We selected 21 V_1CV_2 utterances out of 165 materials mentioned above; V_1=/a/, /e/, /o/, C=/m/, /n/, /g̃/, /b/, /d/, /r/, /z/, and V_2=/a/. We segmented each VCV syllable into following five parts by the observation of speech wave

displayed on CRT connected to MELCOM/70, and stored in a disk file of NEAC 2200/250 again.

A. stationary part of preceding vowel.
B. on-glide.
C. central part of voiced consonant.
D. off-glide.
E. stationary part of following vowel.

Fig. 2. Segmentatation of VCV utterance.

Fig. 2 shows these parts. The criterion of glide taken out of speech wave was such that it contained two or three pitch periods and contained explicitly the influence of both consonant and vowel. Segmented speech waves were compiled as stimulus of following four listening tests: (a) A+C+E, (b) A+C+C+E, (c) A+B+D+E, (d) A+B+S+D+E, where S denotes the silence part of 20 ms. (b) has the longer processing time for perception of voiced consonants than (a). Note that there is no such difference on processing time for machine recognition.

The confusion matrices of these perceptual experiments are shown in Table 4 (a)~(d). The intelligibility of (a)~(d) was 69.5%, 75.9%, 85.7% and 84.8%, respectively. The confusion between /z/ and /d/ was sustained by the similarity between these spectra of this speaker (see Table 10). We should note that the stationary part of vowel might also contain the influence of consonant and there-

Table 4. Confusion matrics of voiced consonants taken out of VCV utterances.

(a) A+C+E (69.5%)

in \ out	m	n	g̃	b	d	r	z
m	100						
n		100					
g̃			93.7			6.7	
b	6.7			86.7		6.7	
d			20		6.7	73.3	
r						100	
z					60	40	

(b) A+C+C+E (75.9%)

in \ out	m	n	g̃	b	d	r	z
m	100						
n	6.7	93.3					
g̃			80			20	
b				93.3		6.7	
d			6.7	6.7	40	46.7	
r						100	
z			6.7	6.7	40	46.7	

(c) A+B+D+E (85.7%)

in \ out	m	n	g̃	b	d	r	z
m	100						
n		100					
g̃			100				
b				100			
d					86.7	13.3	
r						100	
z			20.0		66.7		13.3

(d) A+B+S+D+E (84.8%)

in \ out	m	n	g̃	b	d	r	z
m	100						
n		100					
g̃			100				
b				100			
d					100		
r		6.7			6.7	86.7	
z			13.3		80.0		6.7

fore that if we exchange the consonant part for the same kind of consonant in different context, the intelligibility might decrease. Nevertheless, we could conclude at least following two facts for the design of an automatic recognizer.

(1) More linguistic (or phonemic) information of voiced consonant is contained in glide than in a central part of consonant.

(2) Even if the glide is taken out of a VCV utterance, the intelligibility is beyond 60%~70%.

The second fact is very important knowledge for us. Because we can treat the central part easier than glide. The classification rate, 60%, of voiced consonants is sufficient for an automatic spoken word recognizer or speech understanding system[19].

Next we made an experiment of voiced consonant classification by machine for comparison with human listeners. The speech materials were analyzed by a filter bank, and converted into a series of short time spectra, each of which consisted of 20 components (see the next chapter). We extracted a spectrum from a voiced consonant part for each VCV syllable, and classified this spectrum into one of seven consonants on the basis of Euclidean distance. The reference patterns were calculated from speech materials of ten male adults. Table 5 shows the confusion matrix. This result was better than the perceptual experiment (b). However, if the optimum value is selected as the duration of consonants in the experiment (b), they may become the same result.

Table 5. Confusion matrix of voiced consonant classification by using Euclidean distance.

(79.6%)

in \ out	m	n	g̃	b	d	r	z
m	76	8	4			12	
n	4	76		8	4	8	
g̃		4	60	12		20	4
b		8	4	80	8		
d				7	73		20
r				4		96	
z					4		96

From these experimental results we obtained in conclusion that the precise detection of voiced consonants is more important than the development of the classification technique for automatic voiced consonant recognition.

III Statistical Analysis of Voiced Consonant Spectrum

Statistical analyses of speech spectra have been made by many researchers. For example, techniques of multidimensional scaling and principal component analysis have been devised by a group of Dutch investigators[20),21)]. These men examined the correlation between the physical and perceptual dimensions, using a multidimensional scaling of speech spectra obtained by an 18-channel 1/3-octave

filter bank. Also investigating vowel configuration by use of a principal component analysis, they found a high correlation between the configuration of the average vowels in the factor space and their configuration in the F_1-F_2 formant plane. Other statistical analysis was made by Tabata, who performed multivariate statistics by four factors (speaker, initial vowel, consonant and final vowel) on a set of VCV utterances spoken by five male adults[22].

From different points we investigated the statistical properties of speech spectra. To provide the speech material, 245 meaningful words of VCV-types were spoken by ten male adults. In these words, V was selected from the five vowels /a/, /i/, /u/, /e/, and /o/, and C from the consonants /N/, /y(or j)/, /w/, /m/, /n/, /g/, /b/, /d/, /g/, /r/, and /z/. By visual observation of the spectral patterns of these words, one spectral frame was extracted from each vowel and three from each consonant. We should note, of course, that these are static features of phonemes, although the characteristics of voiced consonants are usually dynamic.

A speech signal is first passed into a pre-emphasis circuit with a slope of 6-dB per octave below 1600 Hz because of improving the signal-to-noise ratio at high frequencies, and then fed into the 20-channel filter-bank. After they are full-wave-rectified and smoothed by the low-pass filter (cut-off frequency: 40 Hz), the output waves are sampled at every 10 ms interval and digitized with an accuracy of 10 bits. The center frequencies of the 20 channels used increase in order by a factor $2^{1/4}$ (210 Hz through 5660 Hz). Since a spectrum is the output of a 20-channel filter bank, it can be considered to be a 20-dimensional vector. Hereafter we will use the following notations for representation of spectrum.

$Z(t)$: the amplitude outputs of the 20-channel 1/4-octave filters at time t, where $Z(t)=z_1(t), z_2(t), \cdots, z_{20}(t)$ are the output of a representation of the output of the sonagraph.

$y(t)$: the normalized $Z(t)$, that is, $y(t)=Z(t)/|Z(t)|$, where $|Z(t)|=(z_1^2(t)+z_2^2(t)+\cdots+z_{20}^2(t))^{1/2}$.

$x(t)$: the logarithmic transformation of $y(t)$, that is, $x_i(t)=\log y_i(t)$.

In speech recognition, we may prefer the relative intensity between frequency components of speech sound spectrum to the instantaneous amplitude $Z(t)$. In other words, we normalize the square sum of the spectral components to 1, yielding $y(t)$. Meanwhile, the auditory sense for the intensity of speech sound is said to be proportional to the logarithm of the intensity itself.

Now we assume that the spectra of each phoneme are distributed in a 20-dimensional vector space according to the multivariate normal distribution $p(x/i)$:

$$p(x/i)=\frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(x-m_i)\Sigma_i^{-1}(x-m_i)^t\right\},$$

where, X is a spectrum, m_i and Σ_i are the mean vector and the covariance matrix of phoneme i, respectively, n is equal to 20 and t denotes the transposition. For every phoneme, we obtained the Σ_i, $|\Sigma_i|$ and m_i for all the speakers and the indivi-

dual speaker.

From these distributions, we can calculate the distance between two phonemes i and j in the spectral space. Various concepts of distance have been dfined for such a purpose, e.g., Minkovski distance, Mahalanobis generalized distance, Kullback distance, Chernoff distance and Bhattacharyya distance, etc. In recently, Goodman[23] calculated the distance between two phonemes from the minimum residual metric used by Itakura[24]. However, this distance is not taken account of the sound variation of the same kind of each phoneme. We use Bhattacharyya distance, because it has a same order relation to the misrecognition rate in the recognition scheme based on Bayes' rule[25].

Bhattacharyya distance B(i, j) is defined as following[26]:

$$B(i,j)=\frac{1}{8}(m_i-m_j)\cdot\left\{\frac{\Sigma_i+\Sigma_j}{2}\right\}^{-1}\cdot(m_i-m_j)^t+\frac{1}{2}\log\left(\frac{|(\Sigma_i+\Sigma_j)/2|}{|\Sigma_i|^{1/2}|\Sigma_j|^{1/2}}\right)$$

Table 6. Bhattacharyya distance between spectral distributions of two phonemes to all speakers.

	a	i	u	e	o	N	y	w	m	n	ŋ	b	d	g	r	z
a	0	8.0	6.1	3.9	3.1	4.4	3.7	2.3	5.7	5.0	4.6	5.9	5.5	6.8	4.0	7.4
i	8.0	0	2.1	3.4	5.5	1.9	2.3	7.5	2.9	2.6	1.3	2.0	2.4	2.2	2.3	3.3
u	6.1	2.1	0	3.3	2.5	1.4	2.8	3.7	1.7	1.6	0.8	0.9	1.7	2.1	1.7	2.8
e	3.9	3.4	3.3	0	3.9	3.3	1.0	4.9	4.5	3.2	2.6	3.5	2.9	4.0	2.1	4.1
o	3.1	5.5	2.5	3.9	0	2.6	3.8	1.1	3.9	3.9	2.5	2.6	3.6	4.4	2.7	5.6
N	4.4	1.9	1.4	3.3	2.6	0	3.2	3.5	1.3	1.3	1.4	1.8	2.7	3.3	2.1	4.5
y	3.7	2.3	2.8	1.0	3.8	3.2	0	4.2	4.0	3.4	2.3	3.0	2.8	4.2	1.6	4.1
w	2.3	7.5	3.7	4.9	1.1	3.5	4.2	0	4.8	4.9	3.7	3.8	4.9	6.3	3.2	7.7
m	5.7	2.9	1.7	4.5	3.9	1.3	4.0	4.8	0	1.0	1.7	1.9	3.1	4.0	2.2	5.6
n	5.0	2.6	1.6	3.2	3.9	1.3	3.4	4.9	1.0	0	1.5	1.7	2.0	3.7	1.6	3.9
ŋ	4.6	1.3	0.8	2.6	2.5	1.4	2.3	3.7	1.7	1.5	0	0.9	1.3	1.6	1.4	2.6
b	5.9	2.0	0.9	3.5	2.6	1.8	3.0	3.8	1.9	1.7	0.9	0	1.2	2.0	1.3	3.5
d	5.5	2.4	1.7	2.9	3.6	2.7	2.8	4.9	3.1	2.0	1.3	1.2	0	1.9	1.5	1.7
g	6.8	2.2	2.1	4.0	4.4	3.3	4.2	6.3	4.0	3.7	1.6	2.0	1.9	0	3.1	2.7
r	4.0	2.3	1.7	2.1	2.7	2.1	1.6	3.2	2.2	1.6	1.4	1.3	1.5	3.1	0	3.5
z	7.4	3.3	2.8	4.1	5.6	4.5	4.1	7.7	5.6	3.9	2.6	3.5	1.7	2.7	3.5	0

Table 6 shows the Bhattacharyya distance, which is calculated by using the spectrum of each phoneme averaged over all the speakers and phoneme environments. From this table, we find that the distance between two vowels is larger than that between two voiced consonants. The distance between two voiced consonants having the same manner of articulation (/m, n, g/ or /b, d, ŋ/) is particularly small. This fact suggests that classification of these phonemes is more difficult than classification of those having the same place of articulation (/m, b/, /n, d/ or /ŋ, g/). Note that the distance B(a, w) between the vowel /a/ and semi-vowel /w/ is comparatively small in spite of the difference in articulation. This is caused by the special fact that the semi-vowel /w/ is always followed by the vowel /a/ in Japanese. Conversely, the distance between /g/ and another phoneme is comparatively large, because /g/ always appears only at the initial position of words.

Table 7. Bhattacharyya distance between spectral distributions of two phonemes for each speaker, averaged over all speakers.

	a	i	u	e	o	N	y	w	m	n	ŋ	b	d	g	r	z
a	0	15.9	11.3	8.4	6.2	12.6	10.4	8.0	13.2	11.5	11.8	10.2	11.0	15.5	7.4	15.9
i	15.9	0	5.4	7.5	13.3	7.3	8.1	21.5	9.9	9.2	5.5	6.0	7.5	7.3	5.8	8.6
u	11.3	5.4	0	6.8	5.6	6.4	8.8	10.6	6.8	6.2	4.5	3.7	5.9	6.0	4.7	7.2
e	8.4	7.5	6.8	0	7.8	10.0	5.7	15.6	12.5	10.0	7.4	8.0	7.9	9.9	5.1	10.8
o	6.2	13.3	5.6	7.8	0	9.4	11.1	6.0	11.3	10.4	8.4	6.5	9.4	10.8	6.4	13.1
N	12.6	7.3	6.4	10.0	9.4	0	17.2	16.8	8.1	7.6	8.1	9.1	12.5	15.5	9.6	17.2
y	10.4	8.1	8.8	5.7	11.1	17.2	0	16.5	18.1	16.0	12.2	11.0	12.1	15.7	8.0	16.1
w	8.0	21.5	10.6	15.6	6.0	16.8	16.5	0	19.2	18.1	14.7	12.1	20.6	20.9	11.6	28.4
m	13.2	9.9	6.8	12.5	11.3	8.1	18.1	19.2	0	6.6	9.0	8.5	13.4	15.1	9.8	20.5
n	11.5	9.2	6.2	10.0	10.4	7.6	16.0	18.1	6.6	0	7.8	7.9	9.5	14.1	8.6	14.3
ŋ	11.8	5.5	4.5	7.4	8.4	8.1	12.2	14.7	9.0	7.8	0	5.6	6.9	9.4	6.7	10.5
b	10.2	6.0	3.7	8.0	6.5	9.1	11.0	12.1	8.5	7.9	5.6	0	5.4	7.1	4.9	9.6
d	11.0	7.5	5.9	7.9	9.4	12.5	12.1	20.6	13.4	9.5	6.9	5.4	0	8.4	6.3	6.8
g	15.5	7.3	6.0	9.9	10.8	15.5	15.7	20.9	15.1	14.1	9.4	7.1	8.4	0	9.4	8.8
r	7.4	5.8	4.7	5.1	6.4	9.6	8.0	11.6	9.8	8.6	6.7	4.9	6.3	9.4	0	9.3
z	15.9	8.6	7.2	10.8	13.1	17.2	16.1	28.4	20.5	14.3	10.5	9.6	6.8	8.8	9.3	0

Table 7 shows the Bhattacharyya distance obtained by averaging the distance calculated for each speaker. Note that it is generally larger than the former in Table 6; however it must be remembered that this is calculated from a smaller number of phoneme samples. If a speaker is fixed, in short, the recognition becomes easier than that for unspecific speakers. This conclusion is consistent with Tabata's conclusion[22], that is, the effect of the speaker-factor is larger than consonant-factor at a stationary part of the nasal consonants.

Table 8 shows the variance of spectral distribution for each phoneme in the spectral space. We use three measures as follows:

(a) The first is a logarithmic transform of the determinant of covariance for each phonemic spectral distribution over all speakers, that is, $\log [\det \Sigma_i]$.

(b) The second is an average over all speakers of the logarithmic transform of the determinant of each phonemic distribution for each speaker (inter-speaker), that is, $\underset{s}{\text{average}} \log [\det \Sigma_i^s]$.

These measures are approximately based on the same order relation of the determinant of volume (or variance) of distribution in the space. For example, we can presume that /a/ is the most stable vowel in the spectral space, and /u/ is unstable.

(c) The third measure is an average of the Bhattacharyya distance over all speakers (inter-speaker), that is, $\underset{j,k}{\text{average}} B(i_j, i_k)$, where i denotes a phoneme, j and k denote speakers.

From third measure, we find that /N/, /y/, /w/, /m/, /n/ and /ŋ/ are more influenced by the speaker-factor than by adjacent phonemes (or phoneme-factor). It is consistent that the nasal spectrum has been effectively incorporated in speaker recognition[27),28)]. If we want to identify a speaker by using vowel spectrum, we had better use the spectrum of /i/.

Table 8. Variance of each phoneme in the spectral space.

(a) logarithmic transform of determinant of spectral distribution to all speakers.

(b) average over all speakers of logarithmic transform of determinant of spectral distribution for each speaker.

(c) average of inter-speaker for Bhattacharyya distance.

	a	b	c	No. of samples
a	-73.4	-84.9	3.9	1088
i	-58.4	-73.0	6.1	838
u	-54.2	-67.4	4.8	880
e	-67.3	-80.9	5.4	885
o	-66.7	-76.6	3.9	921
N	-57.5	-88.3	20.5	750
y	-67.7	-97.3	21.5	450
w	-74.4	-103.2	15.4	450
m	-62.0	-88.8	17.8	750
n	-61.1	-88.4	17.9	750
g̃	-55.6	-77.5	11.9	750
b	-56.9	-73.9	7.8	750
d	-60.0	-82.1	10.5	450
g	-53.3	-71.1	7.2	750
r	-63.5	-80.4	7.1	750
z	-60.1	-80.4	8.8	750

VI Three-Dimensional Representation of Consonants

If each phoneme can be plotted in a few (two or three) dimensional space on the basis of acoustic features, we can understand intuitively the acoustic relationship to one another and get the designing policy of an automatic recognizer. There is the representation by first two or three formant frequencies as the typical model. Klein et al. obtained a three-dimensional representation of vowels by use of a principal component analysis of spectra[21]. Tabata and Sakai plotted Japanese consonants in a three-dimensional space by use of a multivariate statistical analysis of spectra[29]. However, they used the static features of consonants as parameters, although the property of consonants is dynamic. On the other hand, Itahashi et al. tried a three-dimensional representation of Japanese consonants by use of a (non-metric) multidimensional scaling from a viewpoint of perception[30]. They obtained the dissimilarity between phonemes from the confusion matrix of listening test. We also tried a three-dimensional representation of Japanese consonants by use of a (metric) multidimensional scaling like Itahashi et al. But we obtained the dissimilarity from dynamic features of consonants based on their spectra.

First, we must obtain the dissimilarity matrix taking account of dynamic features of consonants. We adopted the time alignment method by using Dynamic Programming. The 58 Japanese monosyllables (CV, C={/m/, /n/, /b/, /d/, /g/, /r/, /z/, /s/, /c/, /h/, /p/, /t/, /k/}, V={/a/, /i/, /u/, /e/, /o/}) were uttered by five

male adults. These utterances were converted into time series of short time spectra by a filter bank. Each spectrum consists of 20 components corresponding to the outputs of a 20-channel filter bank and each spectrum was normalized as 1.

Now, let us consider the dissimilarity D(A, B) between two sequences of short time spectra, $a=a_1, a_2, \cdots, a_I$ and $b=b_1, b_2, \cdots, b_J$, corresponding to two syllables, A and B. We define the distance between two spectra, a_i and b_j, as follows:

$$d(a_i, b_j)=(\sum_{k=1}^{20}(a_{ik}-b_{jk})^2)^{1/2},$$

where $|a|=|b|=1$ and a_{ik} is an output value of k-th filter. Using this distance, we can define the dissimilarity between two syllables as follows:

$$D(A, B)=\min_{(i_1, j_1), \cdots, (i_k, j_k)}[\sum_{k=1}^{k}d(a_{ik}, b_{jk})\cdot \Delta_k]/\sum_{k=1}^{k}\Delta_k,$$

where the sequence of coordinates $(i_1, j_1), (i_2, j_2), \cdots, (i_k, j_k)$ should satisfy following conditions:

$$i_1=j_1=1,\ i_K=I,\ j_K=J$$

$$\begin{cases} i_k=i_{k-1}+1 \\ j_k=j_{k-1}+1 \end{cases} \text{ or } \begin{cases} i_k=i_{k-1} \\ j_k=j_{k-1}+1 \end{cases} \text{ or } \begin{cases} i_k=i_{k-1}+1 \\ j_k=j_{k-1} \end{cases}$$

$$\Delta_k=(i_k-i_{k-1})+(j_k-j_{k-1}), \text{ that is, } \sum_{k=1}^{k}\Delta_k=I+J-1.$$

$$j-r\leq i\leq j+r.$$

Further, the sequence of coordinates can pass continuously to the left to right or bottom to up way only less than m times on the i-j plane, and then must pass to the diagonal way more than n times (constrained DP-matching[31]). We adopted $r=10$ and $m=n=1$. This dissimilarity is calculated easily by using Dynamic Programming technique. This measure of dissimilarity is symmetric and ranges from 0 to 2. Finally, we must define the dissimilarity between two consonants. For example, let us consider the dissimilarity between /m/ and /n/. We define this dissimilarity as following:

$$D(/m/, /n/)=\frac{1}{5}\{D(/ma/, /na/)+D(/mi/, /ni/)+D(/mu, /nu/)$$
$$+D(/me/, /ne/)+D(/mo/, /no/)$$

Table 9 shows the average dissimilarity to five speakers between all pairs of consonants.

As similar, we calculated the dissimilarity between voiced consonants of a male speaker described in Chapter 2. Table 10 shows this dissimilarity matrix. From this matrix, we find that the spectral difference between /z/ and /d/ of this speaker is very small, that is, this fact is consistent with the result of perceptual experiment.

From Table 9, we find the following facts:

(1) The spectral difference between vowels is very large. On the other hand,

Table 9. An average dissimilarity matrix between phonemes for five speakers.

(x1000)

	a	i	u	e	o	y	w	m	n	b	d	g	r	z	s	c	h	p	t	k
a	0	1058	898	896	798	574	422	442	520	516	526	466	458	580	588		360	282	334	318
i	1058	0	784	982	1076			552	558	464		314	456	504	564	390	388	384		316
u	898	784	0	1014	896	554		622	640	508		496	498	700	798	688	516	406		438
e	896	982	1014	0	692			454	396	430	432	432	366	474	552		310	294	334	328
o	798	1076	896	692	0	540		472	532	422	542	408	496	648	698		314	372	462	400
y	574		554		540	0	524	548	554	520	406	500	408	510	612	594	632	550	460	574
w	422					524	0	428	520	426	506	474	430	572	646		474	378	474	500
m	442	552	622	454	472	548	428	0	326	468	472	498	460	606	682	686	560	474	452	546
n	520	558	640	396	532	554	520	326	0	498	410	492	432	550	640	644	600	520	420	538
b	516	464	508	430	422	520	426	468	498	0	374	374	370	512	666	608	498	412	444	476
d	526			432	542	406	506	472	410	374	0	382	346	410	566		554	492	438	492
g	466	314	496	432	408	500	474	498	492	374	382	0	396	466	616	532	474	438	426	384
r	458	456	498	366	496	408	430	460	432	370	346	396	0	480	600	526	486	444	374	476
z	580	504	700	474	648	510	572	606	550	512	410	466	480	0	364	350	602	582	494	540
s	588	564	798	552	698	612	646	682	640	666	566	616	600	364	0	306	620	624	530	548
c		390	688			594		686	644	608		532	526	350	306	0	506	550		440
h	360	388	516	310	314	632	474	560	600	498	554	474	486	602	620	506	0	352	416	360
p	282	384	406	294	372	550	378	474	520	412	492	438	444	582	624	550	352	0	324	400
t	334			334	462	460	474	452	420	444	438	426	374	494	530		416	324	0	342
k	318	316	438	328	400	574	500	546	538	476	492	384	476	540	548	440	360	400	342	0

the difference between /m, n/ and /s, c/, and among /b, d, g, r/ and /h, p, t, k/ is small.

(2) The difference between a vowel and an unvoiced stop is also small. But this is caused by the fact that the duration of an unvoiced stop is much shorter than that of a vowel.

However, we can not know intuitively the spectral relationship among phonemes in the case of a matrix. Now, by using a multidimensional scaling method, we can plot each phoneme into a three-dimensional space on the basis of acoustic features from the dissimilarity between all pairs of phonemes. In general, multi-

Table 10. A dissimilarity matrix between voiced consonants (for Chapter 2).

(x1000)

	m	n	$\tilde{g}$	b	d	r	z
m	0	328	432	354	426	402	422
n	328	0	378	400	336	358	346
$\tilde{g}$	432	378	0	392	300	356	302
b	354	400	392	0	338	350	352
d	426	336	300	338	0	292	238
r	402	358	356	350	292	0	272
z	422	346	302	352	238	272	0

Table 11. Coordinates in the three dimensional space.

	x	y	z
y	-26.7	-140.5	15.8
w	-106.9	44.2	-11.7
m	-157.0	-26.7	-76.4
n	-97.5	-103.6	-70.0
b	-90.3	12.0	89.5
d	-21.0	-89.1	41.2
g	-25.1	15.7	78.8
r	-39.2	-28.6	34.5
z	138.7	-94.8	42.3
s	234.9	-39.8	-79.0
c	217.8	44.2	19.5
h	1.5	177.5	-4.1
p	-57.3	105.2	-14.5
t	-9.8	18.9	-57.6
k	37.8	105.4	-8.1

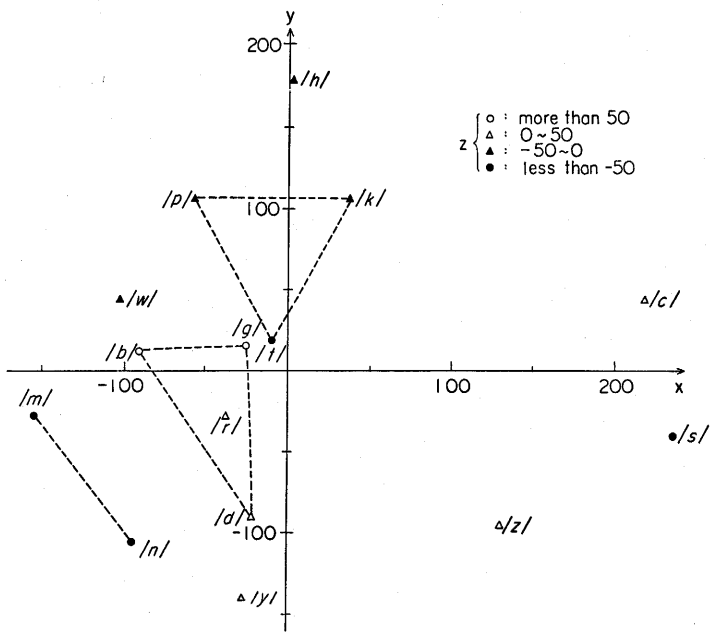

Fig. 3. Three-dimensional representation of Japanese consonants (additive constant=−300).

dimensional scaling methods are divided into two types; metric[32] and non-metric[33]. Since we can regard our definition of dissimilarity as metric measures and it also satisfies approximately the properties of distance, we adopted a metric multidimensional scaling method (see Appexdix). The results are shown in Table 11 and Fig. 3.

From Fig. 3, we can observe the relationship between voiced/unvoiced, mannar of articulation (nasal/stop/fricative/...) and place of articulation (labial/alveolar/palatal/velar...). In particularly, /w/ and /z/ have the property of unvoiced consonants as compared with other voiced consanants.

V Spectral Variation by Speaker-factor

Next, we investigated the spectral difference between an arbitrary pair of five speakers for the same kind of CV syllables similar to Table 8(c). We define the difference of a C_iV_j syllable as following:

$$\frac{1}{{}_5C_2}\sum_{m,n} D(C_iV_j{}^{S_m}, C_iV_j{}^{S_n}),$$

where S_m and S_n denote the kind of speakers. Table 12 shows the results. From these we can conclude following facts:

(1) The spectra of vowels and unvoiced consonants are not so subject to the influence of the speaker-factor as voiced consonants. Therefore the speaker-independent classification of voiced stops is more difficult than unvoiced stops, since the spectral differences among /b/, /d/ and /g/ are almost the same as those

Table 12. Spectral difference between speakers for the same kind of CV syllable.

(x1000)

C \ V	a	i	u	e	o	average
—	520	652	620	454	482	546
y	526		686		452	555
w	406					406
m	532	788	698	550	556	625
n	576	784	720	528	496	621
b	574	762	794	490	542	632
d	534			576	578	563
g	620	710	740	528	544	628
r	506	536	722	520	498	556
z	484	592	628	520	570	559
s	446	460	501	446	352	441
c		514	550			532
h	456	658	702	464	528	562
p	400	670	670	452	426	524
t	500			484	446	477
k	452	672	620	470	488	540

among /p/, /t/ and /k/, and the above mentioned fact.

(2) The spectra of /u/, /i/ and consonants followed by /i/ and /u/ are more subject to the influence of the speaker-factor than others.

However, we should note that those conclusions are not taken account of the spectral variation of a phoneme by many pronunciations of the same speaker in the various contexts unlike the preceding chapter.

Therefore, we investigated the intra-speaker variation for every syllable. The five speakers uttered all syllables on two times and uttered them again after a week and a month. Table 13 shows the ratio of the inter-speaker variation to the intraspeaker variation. From these, we can guess that the most ten efficient syllables for speaker recognition are /i/, /ki/, /ko/, /ga/, /na/, /ni/, /nu/, /hi/, /mi/

Table 13. Ratio of inter-speaker variation to intra-speaker variation.

(a) time interval = immediately

C \ V	a	i	u	e	o	average
—	1.88	2.33	1.96	1.66	1.62	1.89
y	2.27		2.91		1.92	2.37
w	1.51					1.51
m	2.27	2.77	2.42	2.24	2.32	2.40
n	2.47	3.16	2.69	2.28	2.16	2.55
b	2.02	2.27	3.62	1.73	1.91	2.31
d	1.84			1.76	1.60	1.73
g	2.08	1.97	1.98	1.48	1.35	1.77
r	2.06	2.16	2.67	2.08	1.73	2.14
z	1.89	2.06	2.00	1.98	1.83	1.95
s	1.86	1.80	2.37	1.70	1.41	1.83
c		2.42	2.04			2.23
h	1.95	3.29	2.15	1.52	1.93	2.17
p	2.25	2.56	2.48	1.78	1.60	2.13
t	2.29			1.92	1.65	1.95
k	1.92	2.33	2.01	1.77	2.07	2.02

(b) time interval = a week

C \ V	a	i	u	e	o	average
—	1.97	2.22	1.78	1.57	2.19	1.95
y	1.85		2.27		1.47	1.86
w	1.49					1.49
m	1.89	1.98	1.49	1.80	1.65	1.76
n	2.40	2.45	1.81	1.58	1.43	1.93
b	1.89	1.91	1.68	1.19	1.38	1.61
d	2.10			1.85	1.48	1.81
g	2.04	1.66	1.77	1.36	1.53	1.67
r	1.46	1.48	1.56	1.33	1.18	1.40
z	1.57	1.66	1.59	1.54	1.90	1.65
s	1.70	1.43	1.71	1.56	1.40	1.56
c		1.90	1.68			1.79
h	1.78	1.98	1.80	1.40	1.69	1.73
p	1.64	1.80	1.69	1.43	1.37	1.59
t	1.91			1.53	1.74	1.73
k	1.65	2.43	1.50	1.39	2.10	1.81

(c)
time interval = a month

C \ V	a	i	u	e	o	average
—	1.81	1.57	1.53	1.43	1.39	1.55
y	1.88		1.91		1.31	1.70
w	1.35					1.35
m	1.76	1.57	1.55	1.57	1.62	1.61
n	1.91	1.61	1.59	1.48	1.52	1.62
b	1.44	1.76	1.85	1.24	1.31	1.52
d	1.41			1.45	1.43	1.43
g	1.93	1.71	1.88	1.47	1.51	1.70
r	1.59	1.03	1.92	1.39	1.34	1.45
z	1.67	1.64	1.56	1.45	1.90	1.70
s	1.78	1.26	1.50	1.34	1.28	1.43
c		1.54	1.57			1.56
h	1.88	2.02	2.05	1.62	1.67	1.85
p	1.44	1.64	1.51	1.47	1.41	1.49
t	1.91			1.45	1.57	1.64
k	1.81	2.15	1.73	1.69	1.76	1.83

and /yu/ for a short time interval between test and reference samples. However, there is no significant difference of speaker-factor among consonants for a long time interval.

ACKNOWLEDGEMENT

The authors wish to thank Mr. T. Shimamoto and H. Yasuura for their help of perceptual experiments, and also to thank Mr. Y. Miki and T. Tonomura for their cooperation on three dimensional representation of consonants.

REFERENCES

1) T. Sakai and S. Nakagawa: Continuous Speech Understanding System LITHAN, Studia Phonologica IX (1975).
2) T. Sakai and S. Nakagawa: On-line, Real-Time Spoken Words Recognition System with Learning Capability of the Speaker Differences, Studia Phonologica X (1976).
3) K. Nakata and Sugimoto: Processing of speech Information, in *Auditory Sense and Speech,* ed. IECEJ (1968, in Japanese).
4) J. C. R. Licklider and I. Pollack: Effect of differentation, integration, and peak clipping upon intelligibility of speech, JASA, Vol. 20, (1948).
5) T. Sakai, Y. Niimi and K. Ohtani: Various characteristics of speech viewed from computer processing, Tech. Report of the Professional Group on Information Theory and Automata of IECEJ, (1968, in Japanese).
6) J. P. Gupta, S. S. Agrawal and R. Ahmed: Perception of (Hindi) Vowels in Clipped Speech, JASA, Vol. 49, No. 2 (1971).
7) Y. Takeuchi: Perceptual Study of Segmented Japanese Monosyllables, Studia Phonologica I, (1961, in Japanese).
8) S. E. G. Ohman: Perception of Segments of VCCV utterances, JASA, Vol. 40, No. 5 (1966).
9) W. A. Grimm: Perception of Segments of English-Spoken Consonant-Vowel Syllables, JASA, Vol. 40, No. 5 (1966).

10) H. Kuwahara and H. Sakai: Perception of Vowels and C-V Syllables Segmented from Connected Speech, JASJ, Vol. 28, No. 5 (1972, in Japanese).

11) G. A. Miller and P. E. Nicely: An Analysis of Perceptual Confusions Among Some English Consonants, JASA, Vol. 27, No. 2 (1955).

12) S. Saito: Confusion Matrices of Japanese Speech Sounds, Electrical Communication Laboratory Technical Journal (1961, in Japanese).

13) H. Ishigaki, S. Kitayama and S. Soma: Relation between Articulation and Filtering, Reports of the 1976 Autumn Meeting of ASJ (in Japanese).

14) D. W. Marilyn and R. C. Bilgen: Consonant Confusions in Noise: a study of perceptual features, JASA, Vol. 54, No. 5 (1971).

15) C. D. Schatz: The role of context in the perception of stops, Language, Vol. 30 (1954).

16) M. Malecot: Acoustic cues for nasal consonants: An experimental study involving a tape-splicing technique, Language, Vol. 32 (1956).

17) R. A. Cole and B. Scott: The phanton in the phoneme: Invariant cues for stop consonants, Perception and Psychophsics, Vol. 15 (1974).

18) T. Sakai, K. Tabata, H. Ohnishi and S. Kitazawa: Inhouse Computer Network and Host Computer, IPSJ, Vol. 15, No. 12 (1974, in Japanese).

19) T. Sakai and S. Nakagawa: The Effectiveness of Linguistic Information in a Speech Recognition System - An experimental Study by LITHAN-, Paper of Technical Group of Electric Acoustics of IECEJ, Jan. 1977 (in Japanese).

20) L. C. W. Pols, L. J. Tn. Kamp and R. Plomp: Perceptual and Physical Space of Vowel Sounds, JASA, Vol. 46 (1969).

21) W. Klein, R. Plomp and L. C. W. Pols: Vowel Spectra, Vowel Spaces, and Vowel Indentification, JASA, Vol. 48 (1970).

22) K. Tabata and T. Sakai: Multivariate Statistical Analysis of Japanese VCV Utterances, Studia Phonologica VII (1973).

23) R. G. Goodman: Analysis of Language for Man-Machine Voice Communication, Doctor thesis, Standord University, May 1976.

24) F. Itakura: Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Trans. Vol. ASSP-23, Feb. 1975.

25) M. Ichino and K. Hiramatsu: Feature Effectiveness Criteria of Statistical Pattern Classifier, Jour. of IECEJ, Vol. 53-C (1970, in Japanese).

26) A. Bhattacharyya: On a Mearsre of Divergence between Two Statistical Populations Defined by their Probability Distributions, Bull. Calcatt Math. Soc. Vol. 35 (1943).

27) J. W. Glenn and N. Kleiner: Speaker Identification Based on Nasal Phonation, JASA, Vol. 43, No. 2 (1968)

28) J. J. Wolf: Efficient Acoustic Parameters for Speaker Recognition, JASA, Vol. 51, No. 6 (1972).

29) K. Tabata and T. Sakai: Three-Dimensional Representation of Japanese Phonemes, Studia Phonologica VIII (1974).

30) S. Itahashi, T. Kimura and K. Kido: Multidimensional Expression of Japanese Phonemes, Reports of the 1972 Spring Meeting of ASJ (in Japanese).

31) H. Sakoe: Speech Recognition Based on Slope Constrained DP-Matching, Reports of the 1974 Spring meeting of ASJ (in Japanese).

32) W. S. Torgerson: *Theory and method of scaling*, John Wiely & Sons Inc. (1958).
33) J. B. Kruskal: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis, Psychometrica, Vol. 29, No. 1 (1964).

(Aug. 31, 1977, received)

APPENDIX

A Metric Multi-dimensional Scaling Method[32]

Let us denote the experimentally obtained distance between objects i and j by d_{ij}. We suppose that the experimental procedure is inherently symmetrical, so that $d_{ij}=d_{ji}$. We want to represent the n objects by n points in r-dimensional space. We make a matrix from d_{ij} as follows:

$$B=(b_{ij})$$

$$b_{ij}=\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}d_{ij}^2+\frac{1}{n}\sum_{j=1}^{n}d_{ij}^2-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}^2-d_{ij}^2\right)$$

The following theorems from Young and Householder hold for the B matrix.

1. If the matrix B is positive semidefinite, the distances between the stimuli may be considered as distances between points lying in a real, Euclidean space.
2. The rank of positive semidefinite matrix B is equal to the dimensionality of the set of points.
3. The positive semidefinite matrix B may be factored to be obtain a matrix A, where

$$B=AA'$$

If the rank of matrix B is equal to r, where $r\leqq n$, then matrix A is an $n\times r$ rectangular matrix whose elements are the projections of the points on r orthogonal axes with origin at the i-th point of the r-dimensional, real Euclidean space, that is, $\sqrt{\lambda_j}z_{ij}$ represents the coordinate on the j-th axis for the i-th object.

$$a_t=\sqrt{\lambda_t}z_t' \qquad z_t=\begin{pmatrix} z_{1t} \\ z_{2t} \\ \vdots \\ z_{nt} \end{pmatrix} : \text{eigenvector for the eigenvalue } \lambda_t.$$

$$A=(a_1, a_2, \cdots, a_r)=\begin{array}{c} \\ i \end{array}\begin{array}{c} \begin{matrix} \qquad & \qquad & \qquad & j & \qquad \end{matrix} \\ \begin{pmatrix} \sqrt{\lambda_1}z_{11} & \sqrt{\lambda_2}z_{12} & \cdots & \cdot & \cdots & \sqrt{\lambda_r}z_{1r} \\ \sqrt{\lambda_2}z_{21} & \sqrt{\lambda_2}z_{22} & \cdots & \cdot & \cdots & \sqrt{\lambda_r}z_{2r} \\ \vdots & & & \vdots & & \\ \cdot & \cdot & & \sqrt{\lambda_j}z_{ij} & \cdot & \\ \vdots & & & \vdots & & \\ \sqrt{\lambda_1}z_{n1} & \sqrt{\lambda_2}z_{n2} & \cdots & \cdot & \cdots & \sqrt{\lambda_r}z_{nr} \end{pmatrix} \end{array}$$